SUBJECTIVE ASSESSMENT OF GLOBAL PICTURE-WISE JUST NOTICEABLE DIFFERENCE

Hanhe Lin[1]*, Mohsen Jenadeleh*[1]*, Guangan Chen*[1]*, Ulf-Dietrich Reips*[2]*, Raouf Hamzaoui*[3]*, Dietmar Saupe*[1]

[1]Department of Computer and Information Science, University of Konstanz, Germany
[2]Department of Psychology, University of Konstanz, Germany
[3]School of Engineering and Sustainable Development, De Montfort University, UK

ABSTRACT

The picture-wise just noticeable difference (PJND) for a given image and a compression scheme is a statistical quantity giving the smallest distortion that a subject can perceive when the image is compressed with the compression scheme. The PJND is determined with subjective assessment tests for a sample of subjects. We introduce and apply two methods of adjustment where the subject interactively selects the distortion level at the PJND using either a slider or keystrokes. We compare the results and times required to those of the adaptive binary search type approach, in which image pairs with distortions that bracket the PJND are displayed and the difference in distortion levels is reduced until the PJND is identified. For the three methods, two images are compared using the flicker test in which the displayed images alternate at a frequency of 8 Hz. Unlike previous work, our goal is a global one, determining the PJND not only for the original pristine image but also for a sequence of compressed versions. Results for the MCL-JCI dataset show that the PJND measurements based on adjustment are comparable with those of the traditional approach using binary search, yet significantly faster. Moreover, we conducted a crowdsourcing study with side-by-side comparisons and forced choice, which suggests that the flicker test is more sensitive than a side-by-side comparison.

Index Terms— Just noticeable difference, subjective quality assessment, flicker test, paired comparison, JPEG

1. INTRODUCTION

Image compression schemes, e.g., JPEG, are widely used to meet the constraints of transmission bandwidth and storage. With the increase of the compression level, more and more subjects could perceive the image compression artifacts that affect their visual quality of experience. The picture-wise just noticeable difference (PJND) is the smallest distortion that a subject can perceive when an image is compressed. Given the PJNDs of a group of subjects, the fraction of subjects who do not see any distortion when comparing an original image with its compressed version is called the satisfied user ratio (SUR).

Determining the relationship between the compression scheme and user satisfaction is challenging, but essential because of the practical applications. For example, content providers want to model the relationship between bitrate and user satisfaction to maximize user satisfaction with a fixed bitrate. Thus, conducting a subjective study to identify PJNDs of subjects is a prerequisite for studying this relationship.

Many methods may be used to assess the PJND. A baseline method is given by linear or full search. The reference image is compared with the sequence of compressed images with decreasing bitrate until a difference is noticed. This is a linear search, called the "method of limits" in psychophysics. In full search, randomized comparisons with all compressed images are carried out. In the baseline method many unnecessary comparisons may have to be made. Therefore, in previous research, more efficient search strategies based on the bisection method were applied.

- *Standard (aggressive) binary search.* A binary search algorithm can speed up the baseline annotation procedure. The search procedure helps to quickly narrow down the first noticeable difference, resulting in fewer subjective comparisons than the linear search.
- *Relaxed binary search.* This is a modification of standard binary search where the size of the bracketing interval is scaled by 3/4 instead of 1/2 in each iteration. The relaxed version takes longer but is more robust with respect to the nondeterministic outcomes of comparisons.
- *Paired comparisons with scale reconstruction.* Many two-alternative forced-choice (2AFC) comparisons are made, also between compressed images. Subjects identify the image with higher quality. A pair for which one of the images collects 75% of the votes is considered to have a perceptual distance equal to 1 JND.

PJND assessment methods can further be grouped according to the way the images are presented for comparison. The reference and the test image can be displayed sequentially or simultaneously for a specified duration. In the latter case,

978-1-7281-1485-9/20/$31.00

the two images can be viewed side-by-side or on top of each other, alternating at a certain frequency.

For industrial applications, an estimation of the PJND for any input media elements is required without lengthy user studies. Such prediction methods are typically trained with PJND datasets previously acquired in laboratory or crowdsourcing studies. Particularly, for predictions based on deep learning, large datasets are required for training. However, very few such datasets are available, see Section 2. To facilitate an efficient creation of large PJND datasets, we compare several PJND assessment methodologies.

There is no unique definition of a PJND. A just noticeable difference can be detected by observers at differing levels of distortion, depending on the way images are assessed and compared. For example, the results depend on the display size, viewing distance, ambient illumination, and on whether a single or double stimulus method is used. Therefore, the actual PJND values in a study or in an application depend on the choice of the technique for their measurement. For image communication systems, perceptually lossless or near-lossless encoding methods are most relevant. Thus, it is of practical advantage to (1) have a sensitive test for just noticeable differences, and (2) extend these measurements so that one can assign a fractional PJND level for all distorted images, e.g., for images compressed at arbitrary quality factors.

Regarding the desired high sensitivity of PJND tests, the JPEG-XS standard has adopted a flicker test [1]. Instead of comparing a test image side-by-side with a reference image, the reference and compressed images are displayed sequentially, rapidly alternating at a frequency of 8 Hz. In order to achieve (2), the PJND tests need to be extended to a global one so that the PJND for several reference images (i.e., the source image, compressed at different bitrates) are attained. From these results, quality scale values in JND units can be reconstructed, similar to those derived from standard paired comparisons, see Section 3.2 and [2]. Therefore, in this paper, we also apply the flicker test and we assess the PJND for a sequence of (compressed) reference images.

The main contributions of this paper are as follows.

- We introduce a global PJND assessment for a source image where PJNDs are measured according to increasing distortion levels of the reference image.
- We introduce and evaluate two methods of adjustment for PJND assessment, a slider-based method and a keystroke-based method, both using the flicker test.
- We show that the measurements are comparable with those of the traditional binary search, yet significantly faster.
- With a crowdsourcing study, we show that measurements with the flicker test are more sensitive than common side-by-side comparisons with a forced choice.

2. PREVIOUS WORK

In this section, we survey the main subjective quality assessment studies to collect PJND annotations. We also point to the JND-based image and video datasets that have been built in these studies. Later, we discuss papers that are also relevant to the methods used in this contribution.

2.1. Assessment of JND for images

Jin *et al.* [3] conducted subjective quality assessment tests to collect JND samples for JPEG compressed images and built a dataset called MCL-JCI. The tests involved 150 participants and 50 source images of resolution 1920×1080. From each source image, 100 compressed versions were generated by varying the JPEG quality factor (QF) from 1 (lowest) to 100 (highest). The image compressed with QF $= 100$ was used as a reference image. The reference image and a compressed image were displayed side by side on a 65-in TV with a resolution of 3840×2160 to determine whether they are noticeably different. The viewing distance was 2 m from the center of the monitor. For a given image, JND samples were collected from 30 subjects. The standard binary search was used to speed up the process. The study found that humans can distinguish only a few distortion levels (five to seven).

Fan *et al.* [4] studied the JND of symmetrically and asymmetrically compressed stereoscopic images for JPEG2000 and H.265 intra-coding. The study considered 12 stereo images and was conducted by 36 subjects. Stereo image pairs (one source pair and one distorted pair) were shown side by side on a 65-in 3D monitor with native resolution 3840×2160. The subjects wore polarized glasses and were seated 1.6 m from the monitor. Relaxed binary search was used to collect the JND sample from a subject. Outlier subjects were detected, and their PJND samples were removed.

Liu *et al.* [5] created a JND dataset for panoramic images viewed using a head-mounted display. The study involved 42 participants and included 40 source images of resolution 5000×2500. JPEG was used to compress each source image with 100 quality factors. The reference image and a compressed image were displayed simultaneously in random order. For each source image, the test included at least 25 observers. The standard binary search was used to identify the JND. Outliers were removed based on range and standard deviation.

2.2. Assessment of JND for video

Wang *et al.* [6] considered 30 source video sequences of resolution 1902×1080, duration 5 s, and different frame rates. They compressed the video sequences by varying the quantization parameter (QP) of the H.264/AVC video coder from 1 (smallest distortion) to 51 (highest distortion). More than 150 people participated in the study. The viewing distance and display monitor were as in [3]. The video sequence corresponding to QP = 1 was used as a reference. The reference

and a distorted version were displayed one after another. JND samples were collected from 50 subjects. The standard binary search was used to speed up the process. The resulting JND dataset was called MCL-JCV.

The study in [7] involved five source images and five video sequences of resolution 1920×1080. The images were encoded with JPEG, while the video sequences were encoded with H.264 and H.265. The viewing distance and display monitor were as in [3]. The standard binary search was used to speed up the process.

Wang *et al.* [8] built a large-scale JND video dataset called VideoSet for 220 5-s source videos in four resolutions (1080p, 720p, 540p, 360p). Each source video was compressed with H.264 using QP values from 1 to 51. The viewing distance was set according to the ITU-R BT.2022 recommendation. The source video and a distorted version were displayed one after another. A relaxed binary search was used to collect the JND sample from a given subject. At least 30 subjects were involved in the JND estimation of each video sequence. Unreliable subjects and outlying samples were removed.

In [9], 40 HD (1920×1080) source video clips were considered. Each clip had a duration of 5 s and a frame rate of 30 fps. Each clip was compressed with H.265 by varying the QP value from 1 to 51. The source clip and a distorted version were played side-by-side time-synchronously on a 65-in 4K UHD TV. For each source clip, 30 subjects participated in the test. The distance between a subject and the screen was three times the screen height. Standard binary search was used. Outlying samples were excluded with the three-sigma rule.

2.3. Other relevant work

Hoffman and Stolitzka [10] proposed tests to determine if a compressed image differs from a reference image by more than one JND. The testing environment was according to ISO 3664. The monitor used had a 24.3-in diagonal and resolution 1920×1200. A reference (uncompressed) image and an image consisting of the alternating reference image and a distorted version were presented side by side. The observer had to identify which of the two images was non-flickering. A database of about 250,000 responses collected from 35 observers to 18 images was made available. The flicker method proposed in this paper was adopted as a standard [1].

Zhang *et al.* [11] collected a large-scale dataset of perceptual judgements, which included asking subjects whether one reference patch and one distorted patch are identical. They used 20 types of distortions (e.g., photometric distortions, noise, blurring, and compression artifacts) and sequentially composed pairs of distortions. The two patches had a resolution of 64×64 and were shown for 1 s each, with a 250 ms gap in between.

Redi *et al.* [12] compared the performance of absolute category rating obtained by the single stimulus (SS) method with that of the quality ruler (QR) method. The QR consists of a series of reference images varying in a single attribute (sharpness), with known and fixed quality differences between the samples, given by a certain number of JND units [13]. For the QR method, the quality of an input image is compared to the image qualities on the ruler. The study showed that QR scores have narrower confidence intervals than SS scores.

Visual analogue scales, like sliders, for assessing perceived quantities such as length, area, or sensory stimuli like loudness or taste have been studied in psychology and have been shown to be reliable measurement tools [14].

3. METHODS

In this section we describe the proposed PJND estimation using methods of adjustment and the derivation of JND scale values by paired comparisons that we used to estimate the gain in sensitivity due to the flicker display technique.

3.1. Flicker-test-based PJND estimation in the lab

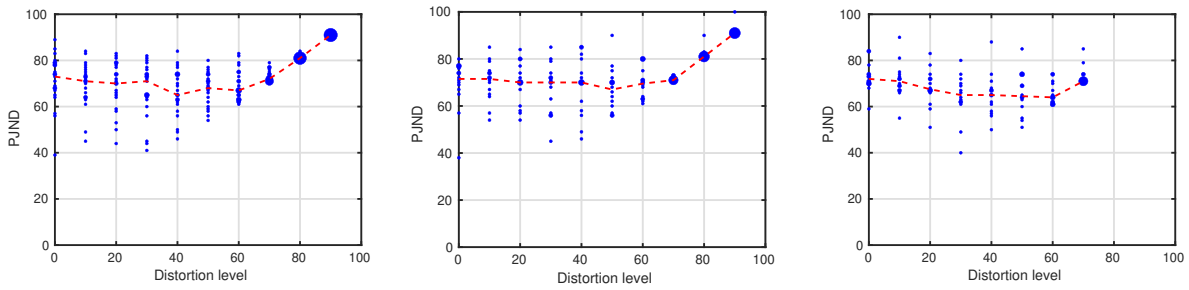
For an image I, we obtain a sequence $I_d, d = 0, 1, \ldots, 100$ with $I_0 = I$ and I_d being the JPEG compressed version with quality factor QF $= 101 - d, d = 0, \ldots, 100$. Thus, as the distortion level d increases, the bitrate decreases. For 10 reference images $I_r, r = 0, 10, \ldots, 90$, our objective is to search for their PJNDs among the images $I_d,\ d > r$.

We used a flicker test, with the reference and the compressed test image being displayed successively at a frequency of 8 Hz. Using this display scheme, we implemented three subjective assessment methods for the PJND.

- *Slider based adjustment.* The marker of a slider controls the distortion level of the test image. Subjects move the slider to the position corresponding to the smallest distortion level with noticeable flicker.
- *Keystroke based adjustment.* The distortion level of the test image is increased or decreased according to key strokes on the left resp. right arrow keys. The initial step size is $\Delta d = 10$, and it is reduced to 5, 2, and 1, each time when the direction is changed.
- *Relaxed binary search.* In this adaptive method, described in Section 1, subjects are only required to determine if a given image pair is flickering or not.

3.2. Crowdsourced PJND estimation: Paired comparison

We also estimated the sensitivity of PJND assessments with plain side-by-side display instead of the flicker test. For this purpose, we applied the method of paired comparisons with scale reconstruction as used, e.g., in the construction of the quality ruler [10]. We selected image pairs (I_k, I_l) with distortion levels k, l from the set $\{0, 10, ..., 90\}$. As this requires a very large number of comparisons, we conducted this study using crowdsourcing.



(a) Slider (b) Keystroke (c) Relaxed binary search

Fig. 1: PJND vs. distortion level of the reference image for source image 48 of the MCL-JCI dataset. Blue disks denote the PJND raw data with areas proportional to the frequencies of the data points. The reference images were derived at distortion levels 0,10,. . .,90. The red-dotted line connects the median values.

From the fractions $A_{k,l}$ of subjects that consider I_k to have a better quality than I_l, we can reconstruct quality scale values on the JND scale using Thurstone's approach (case V) with the maximum likelihood method [15]. From Thurstone's model that assigns Gaussian random variables of equal variance $\sigma^2 = \frac{1}{2}$ to the qualities of all items on the sensory continuum, we conclude that a scale difference of 0.6745 corresponds to 1 JND unit. This is based on a 50% JND, i.e., if two stimuli (I_k, I_l) are 1 JND unit apart, then the detection rate of a just noticeable difference is 50%. In terms of a 2AFC test, this 50% detection rate corresponds to a rate of 75% preference of the item with the better visual quality, since the other half of the observers cannot notice a difference and will be guessing the correct item half of the time.

4. EXPERIMENTS

4.1. Subjective assessment for PJND estimation

We carried out a lab study on selected images from the MCL-JCI dataset [3], see Section 2.1. We sorted all 50 images in ascending order of the mean PJNDs given in the dataset for [3]. The images at positions $5(n-1)+1, n = 1, \ldots 10$ were selected, covering the full range of PJNDs. These 10 images are labeled in the dataset by SRC04, 05, 12, 14, 17, 26, 32, 36, 43, and 48.

The slider- and keystroke-based methods were applied to these 10 images. For the more complex binary search strategy, we reduced to five source images (SRC12, 14, 32, 43, 48) each giving eight reference images: $I_r, r = 0, 10, \ldots, 70$.

We followed the ISO standard [1] to set up the study environment. Before conducting experiments, each participant was informed that the collected data would be processed according to data protection regulation and signed a consent form. Next, each participant filled a form to provide personal information such as age and gender. Finally, each participant was given a set of instructions about the study, including the definition of image flickering and PJND, as well as guidance on how to determine if an image is flickering.

The duration of a study for one image was around 20 minutes. If a participant took part in three studies continuously, there were two 5-minute breaks in-between. Overall, the number of participants was 21 (16 males and 5 females) for the slider-based study; 15 (10 males and 5 females) for the keystroke-based adjustment study; and 14 (9 males and 5 females) for the relaxed binary search study.

4.2. Subjective assessment for paired comparison

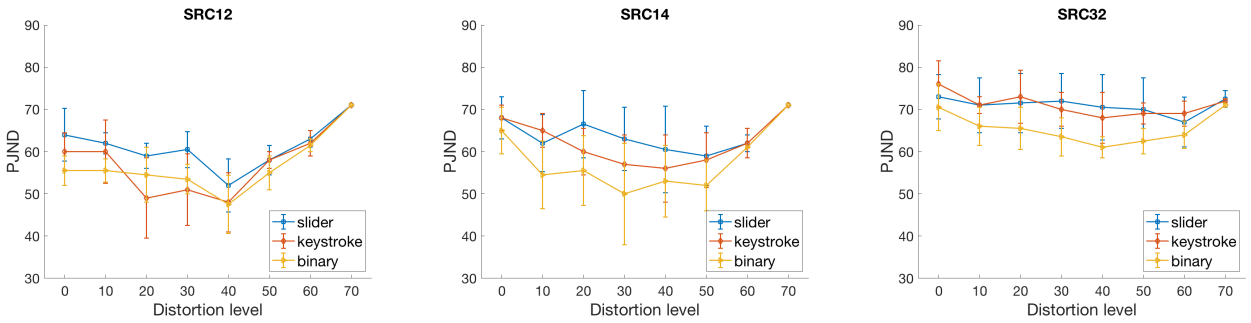
We conducted a crowdsourced paired comparison on the Amazon Mechanical Turk (AMT) platform[1] for five images at eight distortion levels, $I_0, I_{10}, \ldots, I_{70}$ (see Fig. 4). To reduce the image sizes for simultaneous display on common desktop screens, we cropped the images to patches of size 600×480. Given two image patches, crowd workers were asked to identify the one with better quality. Each patch was compared with all others. Each patch pair appeared twice, i.e., an image is on the left side in one pair and on the right in the other one. This resulted in $\binom{8}{2} \times 2 = 56$ pairs with 50 collected votes each. This totaled 2,800 answers per reference image, and 14,000 answers overall, collected from 259 crowd workers.

We removed outliers as follows. First, we discarded all jobs of crowd workers whose number of answers was less than or equal to 10. Second, for each crowd worker, we calculated the true positive rate (TPR) of his/her jobs. The TPR is the fraction of judgements that correctly identified the image of better quality. Here we assumed that the image with the higher bitrate has better quality. We sorted the jobs in ascending order of TPR and iteratively removed the jobs with the lowest TPR until 80% of all answers remained.

5. RESULTS

We analyzed the PJND values for the 10 source images using the three subjective PJND measurement methods (slider-,

[1] www.mturk.com/



(a) MCL-JCI image 12 (b) MCL-JCI image 14 (c) MCL-JCI image 32

Fig. 2: Three examples of PJND medians as functions of distortion level of the reference image, with 95% confidence intervals.

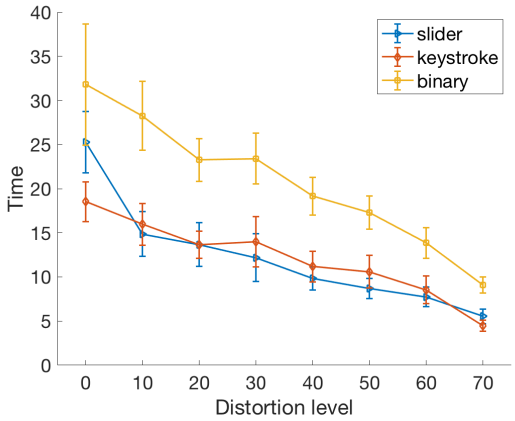


Fig. 3: Time measurements (seconds per PJND) averaged over all participants, with 95% confidence intervals.

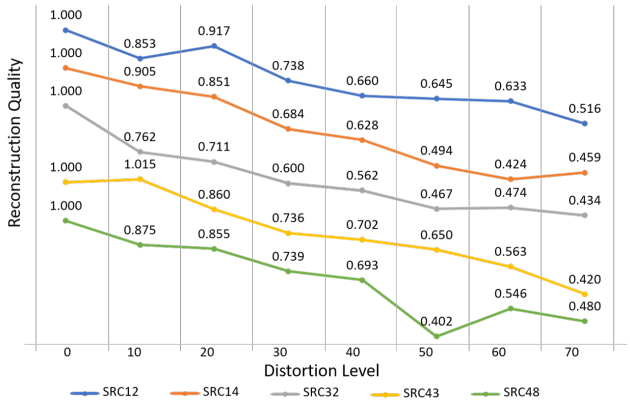


Fig. 4: Image qualities in units of JND obtained by paired comparisons.

keystroke-based, and relaxed binary search). Fig. 1 illustrates an example of the raw data of our PJND measurements. Fig. 2 shows the medians of the PJND values and the corresponding 95% confidence intervals for five selected reference images.

The comparison shows that all three methods produced similar PJND results with generally slightly lower PJNDs for the relaxed binary search. This indicates that the slider- and keystroke-based methods gave comparable results to the relaxed binary search method, with the difference that the relaxed binary search method was more sensitive compared to the adjustment methods. Also it had a higher standard deviation across all levels of distortion.

Fig. 3 compares the response times for the three methods. The average processing time per PJND measurement decreased with the distortion level of the reference image. This is because the number of images that are compared to a reference image decreases with the increase of the distortion level of the reference and because the distortions at consecutive high distortion levels are easily distinguishable. In comparison, the relaxed binary search method takes roughly 1.5 to 2 times more time than the two adjustment methods. With the slider or keystrokes, many redundant comparisons can be skipped rapidly. The average time per JND measurement for the slider, keystroke, and relaxed binary search methods were 12.22, 12.11, and 20.77 s, respectively.

Fig. 4 shows the results of the side-by-side comparisons without flicker in the crowdsourcing study. The reconstructed scale values (in JND units) were shifted so that the original source images have a quality equal to 1 JND. In summary, the average drop in quality from distortion level 0 (no distortion) to level 70 is only about 0.54 JND. In contrast, for the flicker-based tests the PJND for reference level 0 typically is in the range from 60 to 70. In other words, the flicker test provides about twice the sensitivity of a side-by-side comparison.

It should be expected that as the distortion level of the reference image increases, also the PJND increases. Therefore, it is surprising that in Fig. 2 the medians of the PJNDs initially decrease up to distortion level of about 40 and only then begin to increase. This observation can be made for all PJNDs assessed by the three subjective tests across all 10 source images. This finding is in line with Fig. 4: The perceptual image quality does not monotonically decrease. For a deeper understanding of this result, further tests would be required.

6. CONCLUSION AND FUTURE WORK

To estimate the PJND efficiently, we introduced two subjective assessment methods, a slider-based method and a keystroke-based method, and compared them with a traditional one, the relaxed binary search method. We applied the flicker test and assessed the PJND for 10 reference images with different distortion levels.

Compared with the relaxed binary search method, the proposed methods are faster, however, less sensitive. Moreover, an additional crowdsourced paired comparison showed that the flicker test is about twice as sensitive as the classical side-by-side comparisons with forced choice.

To compare our lab-based results using the flicker test with PJND estimates in the crowd, also using the flicker test, we will carry out a pilot study using the full search paradigm.

Future investigations of subjective assessment of global picture-wise just noticeable difference should uncover the response process during the measurement. Understanding this process will help us develop and optimize the dynamic subjective assessment of PJND.

Acknowledgment

This research was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161 and the Exzellenzstrategie des Bundes und der Länder.

7. REFERENCES

[1] ISO/IEC JTC 1/SC 29/WG 1, "29170-2 Information technology – Advanced image coding and evaluation – Part 2: Evaluation procedure for nearly lossless coding," Tech. Rep. JTC 1 / SC 29 N 14760, ISO/IEC, Geneva, February 2015.

[2] R. Luce and E. E. Galanter, "Discrimination," in *Handbook of Mathematical Psychology*, vol. 1, ch. 4, pp. 191–243, John Wiley, 1963.

[3] L. Jin, J. Y. Lin, S. Hu, H. Wang, P. Wang, I. Katsavounidis, A. Aaron, and C.-C. J. Kuo, "Statistical study on perceived JPEG image quality via MCL-JCI dataset construction and analysis," in *IS&T International Symposium on Electronic Imaging 2016*, vol. Image Quality and System Performance XIII, pp. 1–9, 2016.

[4] C. Fan, Y. Zhang, H. Zhang, R. Hamzaoui, and Q. Jiang, "Picture-level just noticeable difference for symmetrically and asymmetrically compressed stereoscopic images: Subjective quality assessment study and datasets," *Journal of Visual Communication and Image Representation*, vol. 62, pp. 140–151, 2019.

[5] X. Liu, Z. Chen, X. Wang, J. Jiang, and S. Kowng, "JND-Pano: Database for just noticeable difference of JPEG compressed panoramic images," in *Pacific Rim Conference on Multimedia (PCM)*, pp. 458–468, Springer, 2018.

[6] H. Wang, W. Gan, S. Hu, J. Y. Lin, L. Jin, L. Song, P. Wang, I. Katsavounidis, A. Aaron, and C.-C. J. Kuo, "MCL-JCV: a JND-based H. 264/AVC video quality assessment dataset," in *IEEE International Conference on Image Processing (ICIP)*, pp. 1509–1513, 2016.

[7] J. Y. Lin, L. Jin, S. Hu, I. Katsavounidis, Z. Li, A. Aaron, and C.-C. J. Kuo, "Experimental design and analysis of jnd test on coded image/video," in *Applications of Digital Image Processing XXXVIII*, vol. 9599, p. 95990Z, International Society for Optics and Photonics, 2015.

[8] H. Wang, I. Katsavounidis, J. Zhou, J. Park, S. Lei, X. Zhou, M.-O. Pun, X. Jin, R. Wang, X. Wang, *et al.*, "VideoSet: A large-scale compressed video quality dataset based on JND measurement," *Journal of Visual Communication and Image Representation*, vol. 46, pp. 292–302, 2017.

[9] Q. Huang, H. Wang, S. C. Lim, H. Y. Kim, S. Y. Jeong, and C.-C. J. Kuo, "Measure and prediction of HEVC perceptually lossy/lossless boundary QP values," in *2017 Data Compression Conference (DCC)*, pp. 42–51, IEEE, 2017.

[10] D. M. Hoffman and D. Stolitzka, "A new standard method of subjective assessment of barely visible image artifacts and a new public database," *Journal of the Society for Information Display*, vol. 22, no. 12, pp. 631–643, 2014.

[11] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

[12] J. Redi, H. Liu, H. Alers, R. Zunino, and I. Heynderickx, "Comparing subjective image quality measurement methods for the creation of public databases," in *The Proceeding of SPIE-IS&T Electronic Imaging, Image Quality and System Performance VII*, vol. 7529, p. 752903, International Society for Optics and Photonics, 2010.

[13] B. W. Keelan and H. Urabe, "ISO 20462: A psychophysical image quality measurement standard," in *The Proceeding of SPIE-IS&T Electronic Imaging, Image Quality and System Performance*, vol. 5294, pp. 181–189, International Society for Optics and Photonics, 2003.

[14] U.-D. Reips and F. Funke, "Interval-level measurement with visual analogue scales in internet-based research: VAS generator," *Behavior Research Methods*, vol. 40, no. 3, pp. 699–704, 2008.

[15] M. Perez-Ortiz and R. K. Mantiuk, "A practical guide and software for analysing pairwise comparison experiments," *arXiv preprint arXiv:1712.03686*, 2017.